A Phenomenological Model of Protein Folding

Ulf H. Danielsson,[1,*] Martin Lundgren,[1,†] and Antti J. Niemi[1,2,3,‡]

[1] *Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*
[2] *Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France*
[3] *Chern Institute of Mathematics, Tianjin 300071, P.R. China*

We construct a phenomenological effective field theory model that describes the universality class of biologically active single-strand proteins. The model allows both for an explicit construction of native state protein conformations, and a dynamical description of protein folding and unfolding processes. The model reveals a connection between homochirality and protein collapse, and enables the theoretical investigation of various other aspects of protein folding even in the case of very long polypeptide chains where other methods are not available.

PACS numbers: 87.15.A- 87.15.Cc 87.14.hm

Various techniques have been developed for the theoretical analysis of protein folding [1], which is maybe the most important problem in molecular biology. In particular all-atom simulations that employ very accurate semi-empirical potential energy functions, facilitate at least in principle a high resolution description of the folding process. But computationally the problem is NP-hard and thus far these very powerful methods have been limited to proteins that have only a relatively small degree of polymerization [1].

Here we present a phenomenological effective field theory model that describes in a very realistic manner the dynamical details of protein folding, even in the case of very long polypeptide chains. Effective field theory models are often employed and sometimes even with great success, to address complicated problems when the exact theoretical principles are either unknown, or have a structure that is too complex for analytic or numerical treatments. Familiar examples of powerful and predictive effective field theory models include the Ginzburg-Landau approach to superconductivity [2] and the Skyrme model of atomic nuclei [3].

In polymer physics field theory techniques became popular after de Gennes [4], [5] showed the equivalence between the self-avoiding random walk and the $N \to 0$ limit of the $O(N)$ symmetric $(\phi^2)^2$ scalar field theory, and proposed that polymer collapse can be modelled by including an additional $(\phi^2)^3$ self-interaction. This approach is very powerful in characterizing the critical properties of polymers. But to our knowledge there are no effective field theory models that detail the dynamics of protein folding and explicitely describe the native state conformations of proteins.

Here we present a phenomenological effective field theory that resides in the same universality class with biologically active proteins. The model enables both the construction of realistic three dimensional protein conformations, and a detailed analysis of the dynamics of folding and unfolding processes. This makes our model particularly valuable tool in the study of phenomenological aspects of protein folding. In particular, our model can describe the details of proteins folding even in the case of very long polypeptide chains where the semi-empirical all-atom simulations still have a long way to go.

The compactness index ν that describes how the radius of gyration R_g scales in the number of central carbons N

$$R_g \;=\; \frac{1}{N}\sqrt{\frac{1}{2}\sum_{i,j}(\mathbf{r}_i-\mathbf{r}_j)^2} \propto\; LN^{\nu} \tag{1}$$

is a universal quantity in the limit of large N [5]. Here $\mathbf{r}_i$ $(i = 1, 2, ..., N)$ are the locations of the central carbons and L is a form factor that characterizes an effective distance between monomers. At high temperatures we expect that ν quite universally approaches the Flory value [5] $\nu \sim 3/5$ that corresponds to the universality class of self-avoiding random walk; Monte-Carlo estimates refine this to $\nu \approx 0.588\ldots$ [6]. On the other hand, for biologically active proteins an analysis of the data in Protein Data Bank [7] yields an estimate $\nu \sim 2/5$ [8], [9]. This is in line with the widely held view [1] that native state proteins are in the $\nu = 1/3$ universality class of compact matter. But to our knowledge

*Electronic address: Ulf.Danielsson@physics.uu.se
†Electronic address: Martin.Lundgren@physics.uu.se
‡Electronic address: Antti.Niemi@physics.uu.se

none of the available theoretical models of *protein folding* has until now been able to accurately describe the $\nu \sim 2/5$ (or $\nu \sim 1/3$) scaling law of biological active proteins.

Here we show that the scaling law of biologically active proteins is computed by the free energy of a discrete version of the two dimensional Abelian Higgs model with an $O(2) \sim U(1)$ symmetric Higgs field, originally introduced to describe superconductivity [2]. The variant considered here was employed in [10] to embed string-like configurations in three dimensional space; the $U(1)$ gauge invariance originates from the requirement that the physical properties of a string must be independent of the choice of local frames in the normal planes. This principle of gauge invariance leads us to an essentially *unique* free energy to describe folding proteins,

$$F \;=\; \sum_{i,j=1}^{N} a_{ij}\left\{1-\cos[\omega_{ij}(\kappa_i-\kappa_j)]\right\} \;+\sum_{i=1}^{N}\left\{b_i\kappa_i^2\tau_i^2+c_i\cdot(\kappa_i^2-\mu_i^2)^2\right\} \;+\; \sum_{i=i}^{N} d_i\tau_i \tag{2}$$

Here $i,j=1,...,N$ label the central carbon atoms along the protein backbone. The variable κ_i corresponds to the gauge invariant *signed* modulus of the Higgs field, and τ_i is the space component of the gauge invariant supercurrent [2]. The first term describes long-distance correlations, it is responsible for the derivative term of the Higgs field in the continuum limit. We have introduced the cosine function to tame excessive fluctuations in κ_i. The middle term describes the interaction between κ_i and τ_i, and the symmetry breaking self-interaction of κ_i. Finally, the last term is a discretized one-dimensional version of the Chern-Simons functional [11]. Its presence provides a simple model for the observed homochirality of biologically active proteins, a positive (negative) parameter d_i gives rise to left-handed (right-handed) chirality [12].

We relate the variables in (2) to the protein backbone geometry as follows: The Higgs field κ_i describes the signed Frenet curvature of the backbone at the site i, and τ_i is the corresponding frame independent Frenet torsion. Once the numerical values of κ_i and τ_i are known, the geometric shape of the backbone in the three dimensional space $\mathbb{R}^3$ is obtained by integrating a discretized version of the Frenet equations [13]. This integration also introduces parameters Δ_i, the average finite lengths of the peptide bonds.

The quantities $a_{ij},\omega_{ij},b_i,c_i,\mu_i,d_i$ are free parameters, and different values of these parameters describe different kind of amino acid structures. For simplicity we consider here only the nearest neighbor interactions

$$a_{ij}=\begin{cases} a\cdot(\delta_{i,i+1}+\delta_{i,i-1}) & (i=2,...,N-1) \\ a \qquad\quad (i=1,j=2) \;\;\&\;\; (i=N-1,j=N) \end{cases} \tag{3}$$

For simplicity we also select all the remaining parameters to be independent of the site index i. Our choice corresponds to a homogeneous protein backbone.

We fold the protein iteratively, by free energy minimization. At each iteration step we first generate a new set of values for the curvature and torsion (κ_i,τ_i) using the Metropolis algorithm [14] with a finite Metropolis temperature-like parameter T_M (that can be related to the actual physical temperature by an appropriate scaling). We then construct a new protein backbone by solving the discrete Frenet equations with a fixed and uniform peptide bond length Δ,

$$|\mathbf{r}(s_i)-\mathbf{r}(s_{i-1})|=\Delta \qquad i=2,...,N. \tag{4}$$

Finally, before accepting the new protein backbone we exclude steric clashes by demanding that the distance between any two central carbon atoms in the new backbone satisfies the bound

$$|\mathbf{r}(s_i)-\mathbf{r}(s_j)|\geq z \qquad \text{for} \quad |i-j|\geq 2. \tag{5}$$

We note that in a native state protein it is quite common for z to acquire values that are of the order of, say, 10% smaller than Δ.

Our simulations start from an initial configuration with $\kappa_i=\tau_i=0$. This corresponds to a straight, untwisted protein backbone. Since the initial Metropolis step is determined randomly, essentially by a thermal fluctuation, our starting point has a large conformational entropy. Consequently we expect that statistically our final conformations cover the landscape of native protein states.

The various parameters in (2) are not fully independent but can be related to each other by diverse scaling transformations and changes of variables. We derive additional restrictions on these parameters by comparing the results of our simulations to the properties of biological proteins. For example, in line with native state proteins we impose the constraint that in a full 2π α-helix turn there are on average about 3.6 central carbons. We have arrived at the final parameter values used in our simulations after an extensive trial-and-error procedure, to obtain results that are as close as possible to the universality class of biologically active proteins. A numerical survey around our chosen parameter values suggests that they are optimal, at least locally in the space of parameters.

We have made extensive numerical simulations using configurations where the number N of central carbon atoms lies in the range $75 \leq N \leq 1,000$. For these configurations we typically arrive at a stable folded state after around 1,000,000 steps. The folding process takes no more than a few tens of seconds in a MacPro desktop computer, even for the large values of N. But in order to ensure the stability of our final configurations we have extended our simulations to 22,000,000 steps. Besides thermal fluctuations, we observe no essential change in the folded structures after the initial 1,000,000 steps which confirms that we have reached a native state.

In Figure 1 we have placed all biologically active single-stranded proteins that can be presently harvested from the Protein Data Bank, with the number N of central carbons in the range of $75 \leq N \leq 1,000$. Using a least square linear fit to the data we find for the compactness index the value $\nu_{PDB} = 0.378 \pm 0.0017$, which is in line with the results previously reported in the literature [8], [9]. In Figure 1 we also show how the compactness index ν in our model depends on N when $75 \leq N \leq 1,000$, using a statistical sample of 80 runs for each value of N. When we apply a least square linear fit to our results we find for the compactness index the estimate $\nu = 0.379 \pm 0.0081$, in excellent agreement with the value obtained from the Protein Data Bank. This excellent agreement confirms that our model does describe the universality class of biologically active proteins. From the data in Figure 1 we find that our model

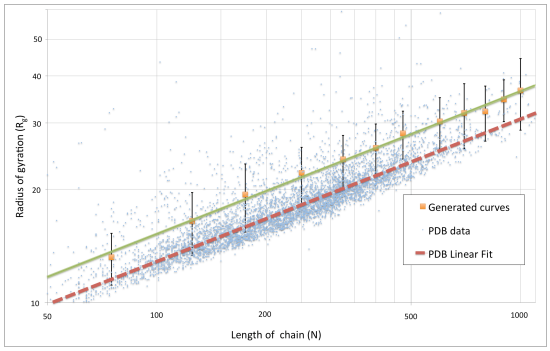

FIG. 1: *Least square linear fit to the compactness index ν computed in our model ($\nu = 0.379 \pm 0.0081$) compared with that describing all single strand proteins currently deposited at the Protein Data Bank ($\nu_{PDB} = 0.378 \pm 0.0017$). The error-bars describe standard deviation from the average, a measure of conformational entropy in our initial configuration.*

predicts for the form factor L in (1) the numerical value $L = 2.656 \pm 0.049$ (Å). This compares well with the average value $L_{PDB} = 2.254 \pm 0.021$ (Å) that we obtain using a least square fit to the Protein Data Bank data displayed in Figure 1.

We observe from Figure 1 that the standard deviation displayed by our final conformations are comparable in size to the actual spreading of biologically active proteins around their experimentally determined average values. The standard deviation is a measure of conformational entropy, and consequently at each value of N our initial configuration appears to have enough conformational entropy for our model to cover the entire landscape of native state protein folds.

We have verified that our value of ν is temperature independent for a wide range of temperatures: In our model ν is insensitive to an increase in the Metropolis temperature T_M until T_M reaches a critical value that we can normalize to $T_c \approx 330\ (K)$. At this critical temperature there is an onset of a transition towards the Θ-point, and at the Θ-point we estimate $\nu \approx 0.48 - 0.49$ in line with the expected value $\nu \sim 1/2$ that characterizes the universality class of a random coil. In the limit of high temperatures we find $\nu \approx 0.65$ which is slightly above the Flory value $\nu = 3/5$ for a self-avoiding random walk. It appears to us that the slight differences between our estimates and those expected on general grounds [5], are due to finite length effects.

We have also studied the effect of the various operators in (2) in determining the universality class:

We find that the value $\nu \approx 0.379\ldots$ is *entirely* due to the presence of the chirality breaking Chern-Simons term: When we remove the Chern-Simons term by setting $d_i = 0$ in (2) the compactness index increases to $\nu \approx 0.488\ldots$ which is very close to the Θ-point value $\nu \sim 1/2$. Thus our model proposes that the folding of biologically active proteins is driven by their chirality.

When we in addition remove the direct coupling between torsion and curvature by setting $b_i = d_i = 0$ the compactness index remains near its Θ-point value $\nu \approx 0.488\ldots$.

When we remove the entire symmetry breaking potential by setting $c_i = 0$ in (2) we find that ν approaches the value $\nu \approx 0.737\ldots$. This proposes that we may have a novel universality class in the low temperature phase. Notice that in this case the local minima of the potential energy are absent.

When we set $a_{ij} = 0$ we find $\nu \approx 0.370$.... Consequently the non-local coupling between curvatures appears to have a tendency to (*very* slightly) increase ν. We also find that in the absence of a_{ij} the value of L tends to increase

slightly, to $L \approx 2.96\ldots$.

Finally, at very high temperatures T_M, when we set $b_i = d_i = 0$ we find that the compactness index, as expected, approaches the Flory value 3/5; we now get $\nu \approx 0.61\ldots$.

In Figure 2 we show using an example with $N = 300$, how the compactness index ν evolves as a function of the number of iterations ("time"), during the first 1,000,000 steps. In this figure we also describe how the free energy (2) develops as a function of the iteration steps. We find that while ν generically approaches its asymptotic value

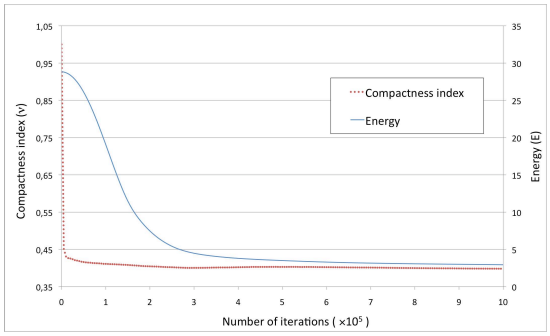

FIG. 2: *The red line shows how the compactness index ν typically evolves as a function of the number of interation steps (time) when computed as an average over statistical samples and up to 1,000,000 steps. The blue line shows similarly how the average energy typically evolves as a function of the number of interation steps.*

$\nu \approx 0.379\ldots$ very rapidly, after only a few thousand iterations, the process of energy minimization typically takes about two orders of magnitude longer. The asymptotic behaviour of the curves confirms that the final state is highly (meta)stable. The stability is further validated by a comparison with Figure 1 where we report on results after the iteration process has been continued by 21,000,000 additional steps: For $T_M < T_\Theta \approx 330\ (K)$ we find no essential change in the final conformations after $N \sim 1,000,000$ steps, beyond thermal fluctuations.

Our interpretation of the Figure 2 is that the folding process described by our model follows very closely the folding process of biological proteins: The initial denatured state first rapidly collapses into a molten globule, with a large decrease in conformational entropy but only a very small change in the internal energy. After the initial collapse to the molten globule with the ensuing formation of secondary structures such as α-helices and β-sheets, the folding process continues with a relatively slow conformational re-arrangement towards a locally stable conformation. The final state has a substantially lower energy than the corresponding molten globule state.

Finally, we have compared our native states to the hierarchical classification scheme CATH [15]. We find that our conformations are in very good overall correspondence with this classification scheme. In particular, our model appears to produce all the major secondary structures of biologically active proteins. But when we compare the number of conformations that we produce in different classes with the data presently available in the Protein Data Bank, we conclude that *as it stands* our model produces a statistical excess of native folds in the mainly-α class in comparison to the data currently deposited in the Protein Data Bank. In order to produce a statistically larger proportion of folds in the α-β and in particular in the mainly-β class, we presumably need to consider a more involved nonlocal coupling (3).

In summary, we have developed a phenomenological effective field theory model that describes realistically the folding dynamics and native state folds of protein backbones. Since our model folds even very long proteins within a few tens of seconds in a desktop personal computer, it allows us to describe and analyze phenomena that are not yet reachable by other theoretical means. The model computes accurately the compactness index ν of native state proteins and in particular it proposes that protein collapse is driven by homochirality which is described by the Chern-Simons functional. Furthermore, since the native states are in line with the CATH classification scheme, the model has great promise to become an effective tool for producing conformational templates for the purpose of protein design and engineering.

Our research is supported by grants from the Swedish Research Council (VR). The work by A.J.N is also supported by the Project Grant ANR NT05-142856. A.J.N. thanks H. Orland for discussions and advice. We all thank M. Chernodub for discussions, and N. Johansson and J. Minahan for comments. A.J.N. also thanks T. Gregory Dewey

for communications. A.J.N. thanks the Aspen Center for Physics for hospitality during this work.

---

[1] L. Mirny, E. Shakhnovich, Annual Review of Biophysics and Biomolecular Structure **30** (2001) 361; H.A. Scheraga, M. Khalili and A. Liwo, Annual Review of Physical Chemistry **58** (2007) 57; M. Oliveberg and P.G. Wolynes, Quarterly Reviews of Biophysics **38** (2005) 245

[2] P.G. De Gennes, *Superconductivity of Metals and Alloys* (Westfield Press, New York 1995)

[3] I. Zahed and G.E. Brown, Physics Reports **142** (1986) 1; T. Gisiger and M.B. Paranjape, Physics Reports **306** (1998) 109

[4] P.G. De Gennes, Physics Letters **38A** (1972) 339

[5] P.G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979)

[6] B. Li, N. Madras and A. Sokal, Journal of Statistical Physics **80** (1995) 661; N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhauser, Berlin, 1996)

[7] H.M. Berman *et.al.*, Nucleic Acids Research **28** (2000) 235

[8] L. Hong and J. Lei, http://arxiv.org/abs/0711.3679v1

[9] T.G. Dewey, Journal of Chemical Physics **98** (1993) 2250

[10] A.J. Niemi, Physical Review **D67** (2003) 106004

[11] S.-S. Chern, J. Simons, Annals of Mathematics **99** (1974) 48

[12] For each i the potential energy in (2) is unbounded from below with a global minimum at $\kappa = 0$ and $\tau \to \infty$. But whenever $b\tau^2 < 2c\mu^2$ there are two symmetric local minima $\mathcal{C}^{\pm}$ in the vicinity of $\kappa \approx \pm\mu$, $\tau \approx -d/2b\mu^2$. The folded proteins are soliton-like configurations that interpolate between these two local minima (at finite temperature).

[13] M. Spivak, *A Comprehensive Introduction to Differential Geometry Volume Two* (Publish or Perish, Inc.. Houston 1999)

[14] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Journal of Chemical Physics **21** (1053) 1087

[15] A.L. Cuff *et.al.*, Nucleic Acids Research (Advance Access published on November 7, 2008)